

# Learning selectivity and invariance through spatiotemporal Hebbian plasticity in a hierarchical neural network

Minjoon Kouh\*

*Physics Department, Drew University, Madison, NJ 07940*

When an object moves smoothly across a field of view, the identity of the object is unchanged, but the activation pattern of the photoreceptors on the retina changes drastically. One of the major computational roles of our visual system is to manage selectivity for different objects and tolerance to such identity-preserving transformations as translations or rotations. This study demonstrates that a hierarchical neural network, whose synaptic connectivities are learned competitively with Hebbian plasticity operating within a local spatiotemporal pooling range, is capable of gradually achieving feature selectivity and transformation tolerance, so that the top level neurons carry higher mutual information about object categories than a single-level neural network. Furthermore, when genetic algorithm is applied to search for a network architecture that maximizes transformation-invariant object recognition performance, in conjunction with the associative learning algorithm, it is found that deep networks outperform shallower ones.

## INTRODUCTION

Object recognition can be considered as a task of determining whether a set of conditions that make up an “object-ness” is satisfied or not. From the lock-and-key mechanism of molecular receptors, to a computer algorithm for face identification, recognition process involves: selecting useful features, learning and memorizing the presence or absence of those features in different object categories, and then classifying an object in question. All recognition systems are also subject to energy, time, and other practical constraints through which their effectiveness and robustness (e.g., quickly and correctly identifying a prey or a predator) are tested.

It is widely acknowledged that the primate visual cortex is a highly efficient recognition system, whose general level of performance is not yet surpassed by engineered systems, as we have not yet “solved” the recognition problem [1]. Object recognition is a computationally difficult task, because it requires for the recognition system to be simultaneously selective for the differences between objects and tolerant to the changes when the same object undergoes different transformations. For example, two different sets of photoreceptors on retina would be activated when the same face appears on the left and the right sides of the visual field, while an overlapping set of photoreceptors are activated when different faces appear on the same side of the visual field. Hence, the visual cortex is charged with producing the same output even when the inputs are very different, as in the former case (i.e., tolerance). At the same time, it must generate different outputs even when the inputs are similar, as in the latter case (i.e., selectivity). A similar mathematical operation would be modulo function, where  $\text{mod}_2(2) = \text{mod}_2(4) \neq \text{mod}_2(3)$ , even though 2 is closer to 3 than to 4.

As demonstrated by a number of biologically-plausible models of visual object recognition [2–5], a hierarchi-

cal processing is one possible computational strategy to satisfy these two requirements for selectivity and tolerance. In these models, the neural population at the upper level becomes more tolerant to object transformations, by combining the responses from less tolerant neural population at the lower levels. At the same time, the neural population becomes selective for different visual features by competitively learning to transmit as much information as possible about various stimuli [6–9].

Hebbian plasticity, which postulates that correlated activities facilitate the strengthening of neural connectivity [10, 11], is a plausible learning mechanism for selectivity and tolerance. As all objects in the world follow the laws of physics, there are predictable structure and inherent coherence in their appearances, which can be exploited by the Hebbian, associative plasticity to acquire neural response properties that are useful for object recognition [6, 8, 12]. Transformation tolerance can be achieved by learning temporal correlations, as an object tends to undergo its transformations smoothly in time [9, 13, 14].

Neuroanatomy and physiological studies of the primate cortex suggest that the visual information is processed hierarchically, so that the receptive fields of the visual neurons in the higher cortical areas are larger, their tuning properties are more complex, and their responses are more tolerant to translation, rotation, and scaling of a stimulus [1]. A potential rationale for such a cortical organization is that constructing a highly selective yet simultaneously tolerant feature detector is computationally difficult, because a highly specific description for an object (e.g., a person with a red jacket) would not be effective in identifying the same object under different circumstances (e.g., the same person with a blue jacket). It has been observed that there is indeed a trade-off between selectivity and tolerance properties [15, 16], and the hierarchical organization seems to be a strategy employed by the neural systems to deal with such an inherent trade-off. Although the trade-off of selectivity and tolerance is

inevitable for a given level in the hierarchy, it is possible to gradually increase the capacity for generalization after several stages of neural computations.

This study corroborates these intuitive ideas, and demonstrates that unsupervised learning of selectivity and tolerance can be performed with a canonical algorithm, where the same plasticity rule is applied repeatedly within a hierarchy. Other recent “deep-learning” approaches [3, 5, 17–19] have also explored similar one-algorithm principle for configuring a recognition system, which is a theoretically attractive hypothesis for its simplicity and an evolutionarily practical solution for building a robust recognition system. Contribution of this study is in the integration of spatial and temporal learning, where temporal correlations in the stimuli endow transformation tolerance [9, 13, 14, 20] and spatial correlations endow selectivity to diverse visual features.

### ALGORITHM

There are three major components in the learning algorithm: Hebbian plasticity, competition, and hierarchical architecture. The schematic diagram of the underlying neural network model is shown in Fig. 1.

1. The Hebbian rule takes advantage of the consistency and persistency of object features in space and time. A synaptic connection is created, if a presynaptic neuron and a postsynaptic neuron are co-active within a predetermined spatial or temporal afferent pooling ranges.
2. Competition allows for the neural network to acquire diverse response properties. During the learning process, a single neuron among a set of neurons with the same afferent pooling range is selected to undergo Hebbian plasticity. The other neurons are suppressed, until a different neuron is selected. The competition is based on the afferent activities. As a result, different neurons learn to specialize in representing diverse patterns of afferent activities. Once the learning phase is complete, competition is turned off.
3. Above two algorithmic elements are embedded hierarchically, so that Hebbian learning of selectivity and tolerance starts at the lowest level and builds up sequentially and gradually to the higher levels.

In the implementation of the algorithm, binary neural responses were considered. The activity of a postsynaptic neuron  $y_i$  is determined by a simple thresholding operation after summing its afferent activities  $x_j$ , which mimics the firing of an action potential when the cumulative influence of the presynaptic input brings the membrane

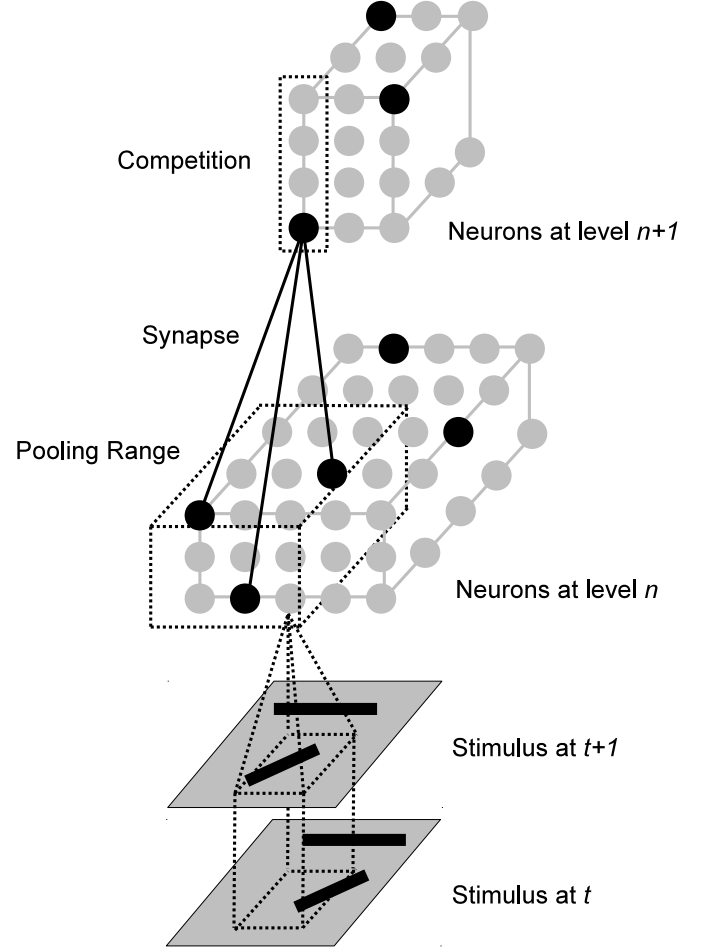


FIG. 1. Schematic diagram of the hierarchical neural network model. Each postsynaptic neuron has a limited spatial and temporal pooling range of presynaptic neurons. There is a competition among the neurons with the same pooling range, and the winning neuron forms the synapses with the active presynaptic neurons within the pooling range. Inactive neurons are depicted with gray circles, and active ones are black. The same rule of synaptic plasticity operates throughout all the levels in the hierarchy, so that the neurons in the upper levels have larger and more complex receptive fields with a bigger range of tolerance to object transformations. Only two levels ( $n$  and  $n + 1$ ) are shown here.

potential above a spiking threshold.

$$y_i = \begin{cases} 1, & \text{if } \sum_j w_{ij}x_j > \theta \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The threshold  $\theta$  determines the sparsity of response [7], and in this study, it was fixed at a constant value of 0.3. A more sophisticated model of neural activity could involve

a microcircuit with divisive normalization for gaussian-like selectivity and maximum-like invariance operations [5].

Synaptic weights were also binary, so that a pair of presynaptic and postsynaptic neurons could be either connected or not. The synapses are initialized randomly. During the learning phase, a synapse is created between a presynaptic neuron  $x_j$  and a winning postsynaptic neuron  $y_i$ . The winning neuron is chosen as the one with the largest presynaptic activity among a pool of neurons having the same spatial and temporal receptive field or belonging to the same cortical column [18, 21, 22], even if its input has reached the threshold  $\theta$ . When there are multiple neurons with the same presynaptic activity, one is randomly chosen. For a winning  $y_i$ ,

$$w_{ij} = \begin{cases} 1, & \text{if } x_j = 1, \text{ where } j \in \mathcal{P} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The pooling range,  $\mathcal{P}$ , is a local spatiotemporal window from which presynaptic responses can be pooled. Such local pooling is crucial for gradually building up selectivity and tolerance, as well as for creating a well-connected network with minimal amount of wiring. Local temporal pooling is consistent with the idea of trace learning [9, 13] and spike-timing-dependent plasticity (STDP) [11, 23]. Two different types of pooling ranges are implemented as shown in Table . One type has a spatially large, temporally short range, so that the postsynaptic neurons develop representation for different spatial features. The other type has a spatially small, temporally long range, so that the temporal sequence of an object transformation is learned. These two types are interleaved hierarchically, so that the neural population at the lowest level learns spatially, and the next level, temporally, and the next level, spatially, and so forth, in a similar fashion as the alternating layers of simple and complex cells in the feedforward, hierarchical models of visual cortex [2, 4, 5].

Another important implementation detail is a mechanism to mark the beginning of each learning cycle, when a new round of competition begins. The onset of an eye movement or phases of various brain waves may serve this role of resetting the learning cycle, while each eye fixation, lasting a few hundred milliseconds on average, is long enough to capture an object going through a short sequence of motion or other transformations.

## RESULTS

### Learning from simple stimuli

The basic operation of the learning algorithm is demonstrated with images of two toy “objects,” where each object is composed of a specific, co-occurring pair of simpler shape elements, namely line segments of four

Simulation parameters		
Level	Pooling Range	Number of Neurons
0	-	$36 \times 36 \times 1$
1	$6 \times 6 \times 1$	$16 \times 16 \times 8$
2	$2 \times 2 \times 4$	$8 \times 8 \times 4$
3	$6 \times 6 \times 1$	$2 \times 2 \times 8$
4	$2 \times 2 \times 4$	$1 \times 1 \times 8$

TABLE I. The zeroth level is for the stimulus image, whose dimension is  $36 \times 36$  pixels. 2500 (500 exemplars at 5 different positions) images were used for learning, and 1000 (200 new exemplars) images were used for calculating the mutual information in Fig. 3. The pooling parameters are listed as horizontal  $\times$  vertical  $\times$  temporal ranges. The number of neurons in each level is determined by the original stimulus size, pooling range, and the degree of overlap between the neighboring pooling ranges. The last number indicates the number of neurons having the same pooling range.

different orientations. The first object was composed of “—” and “\”, and the second, “|” and “/”. Each shape element appeared at non-overlapping random positions, and was translated smoothly in random directions, as shown in Fig. 2(a). As a simplistic analogy, consider two different people: one of them always wearing a blue hat and a green shirt, and the other wearing orange pants and red shoes. A visual recognition system could learn to distinguish these two people by forming a representation about the four different clothing items and their combinations.

When the hierarchical neural network with the proposed learning algorithm is presented with multiple exemplars of these two objects, the lowest level neurons, having the smallest receptive fields, learn the fragments of the four shape elements, as shown in Fig. 2(b). Since these shape elements appear at different positions, the neurons in the next level pool temporally and develop synaptic connectivity with the presynaptic neurons that have the same orientation selectivity at different positions, therefore acquiring tolerance to translation, as shown in Fig. 2(c). This is the same two-level feedforward model of simple and complex cells in the primary visual cortex, constructed through Hebbian plasticity [24]. The neurons in the upper levels with larger receptive fields learn to associate the co-occurring shape elements and form a representation for the two objects, as shown in Fig. 2(d). Such connectivity is also consistent with the bimodal orientation tuning of the intermediate cortical areas, such as V2 and V4 [25–28]. Furthermore, associative object representation between two subsequent primate temporal areas may arise from the same hierarchical learning [29].

The trend of gradual and systematic increase in selectivity and tolerance can also be analyzed with mutual information between neural response  $r$  and stimulus  $s$ ,  $I(r, s)$ , which captures the amount of uncertainty about

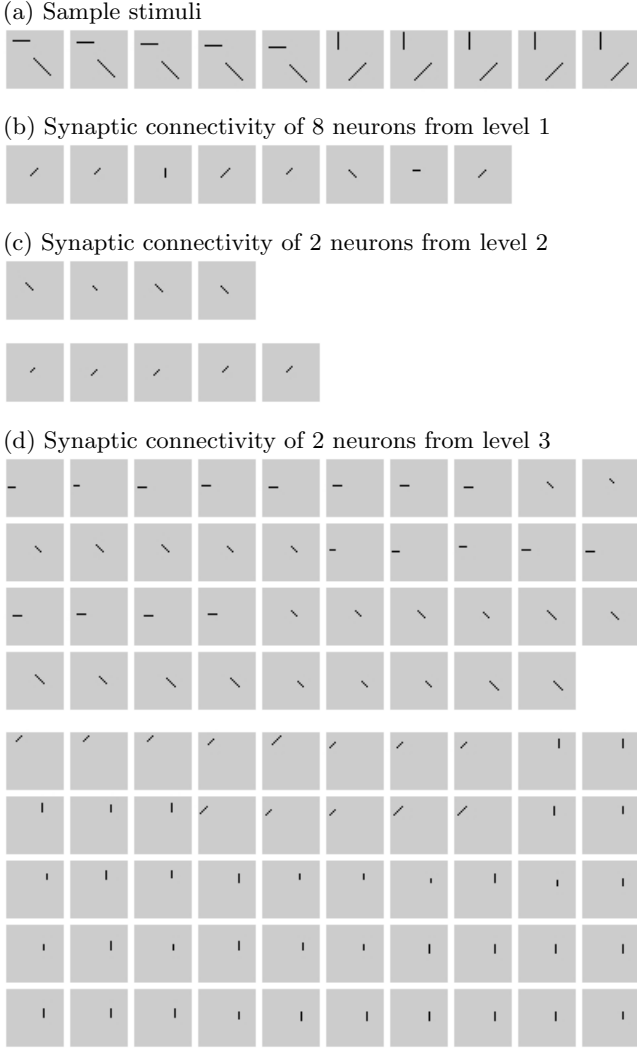


FIG. 2. The spatiotemporal Hebbian learning achieves selective and invariant representation. (a) Ten frames of sample stimulus images. The first five frames show the first object, composed of a specific, co-occurring pair of oriented lines, and the next five frames show the second object. Both objects undergo smooth translations. Each pixel can be considered as the activity of the photoreceptors on the retina. (b) Each panel corresponds to the learned connectivity with the photoreceptors for each neuron in level 1. Hence, each neuron becomes active when the stimulus matches the pattern of synaptic connectivity. (c) The first sample neuron connects with four presynaptic neurons with the same orientation selectivity at different positions. The second sample neuron connects with five presynaptic neurons. These two neurons in level 2 are orientation-selective and translation-tolerant. (d) The first sample neuron pools from the neurons in level 2, which in turn pool from 39 neurons in level 1 overall. This neuron shows the selectivity for one of the objects (i.e., a combination of “—” and “\”) and has a large receptive field. The second sample neuron is selective for the other object. The neurons in the next level (not shown) pool from even more presynaptic neuron, while maintaining consistent object selectivity.

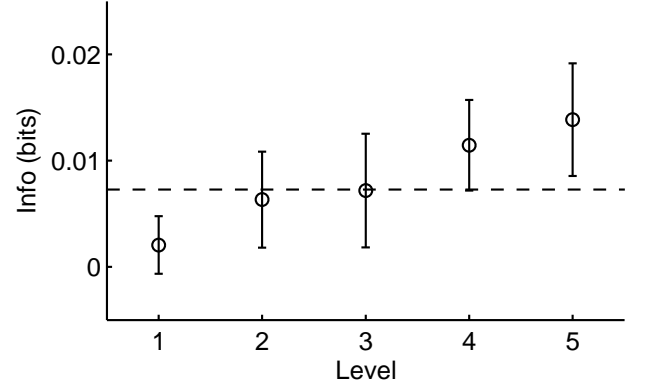


FIG. 3. Mutual information between neural response and stimuli. After synaptic connectivities have been learned from a set of training images, response to a new set images was used to evaluate mutual information. The dotted horizontal line indicates the mutual information of a single-level neural network model, learned with linear regression.

$s$  removed by knowing  $r$  [20].

$$I(r, s) = \sum_{r,s} P(r, s) \log_2 \frac{P(r, s)}{P(r)P(s)}. \quad (3)$$

Since individual shape elements could appear at various positions, the response of the neurons in the lower levels, which are not yet tolerant to translation nor selective for the specific combinations of line segments, would yield low information, while the response of the neurons in the higher levels would yield larger information about the stimulus category. For example, if the response of a neuron is  $r = \{0, 0, 1, 1\}$ , when two exemplars of one object  $O_1$  is shown, followed by two exemplars of different object  $O_2$  (i.e.,  $s = \{O_1, O_1, O_2, O_2\}$ ), the mutual information is 1 bit, since an unambiguous object recognition can be performed with the neural response  $r$ . If the neural responses to the same sequence of stimuli were  $r = \{1, 0, 1, 0\}$ , the mutual information between the stimuli and the response would be zero.

As shown in Fig. 3 and as expected from Fig. 2, the responses of the neural population in the higher levels on average carry more information about the objects. A similar trend is found when object classification performance was compared, using the neural responses from a higher (e.g., inferotemporal cortex) and lower (e.g., V4) visual cortical areas [30].

The most important advantage of such a hierarchical construction of selectivity and tolerance is its generalization capacity. When a single-level neural network is constructed with linear regression and a threshold operation using the same training images as before, its mutual information to the same test images is lower than the average mutual information of the hierarchical model, as shown in Fig. 3, even when the weights of the single-level model were allowed to be non-binary real values.

The threshold was chosen so that the probability of response was equal to the average response probability of the top level neurons in the hierarchical model, which was about 0.01. In other words, a sparsity-matched, single-level neural network carries less information than a multi-level neural network, because the inputs are nonlinearly transformed.

### Learning from natural images and optimizing the architecture with genetic algorithm

Going beyond learning from simple stimuli, a set of gray-scale natural images, collected from the internet, have been applied to the proposed learning algorithm. Each image was presented over several time steps during which it was translated along a randomly chosen direction, simulating a temporally coherent motion. After learning from 250 images, object recognition performance of the learned neural network was measured with the NIST database of handwritten digits. For biological plausibility and simplicity, classification was based on the nearest neighbor scheme of how close the population response vector from the top level was to the average vectors of 10 different categories (digits 0, 1, ..., 9).

The nearest neighbor classifier was trained on 500 images of centered handwritten digits, and the classification performance was measured with 100 test images that were different from the training images and translated by 3 pixels (corresponding to approximately 10 percent of the dimension of the digits, whose sizes were 28x28 pixels) from the center. Then, a genetic algorithm [31] was used to search for an optimal network architecture, using the test performance as a fitness function. This approach was motivated by the idea that the global network architecture may be optimized over a longer, evolutionary time scale, while local features are learned over a shorter time scale through visual experiences. For example, the degree of expression of neural growth factors could determine the size of dendritic arbor [32] and spatiotemporal pooling ranges, which would determine the number of hierarchical layers necessary to cover a desired field of view. Furthermore, the recognition performance on the transformed test images would be a more relevant fitness function for the goal of building a selective and invariant visual system.

In this particular implementation of genetic algorithm, six parameters in the proposed learning algorithm, have been expressed as a set of binary numbers, which would go through crossover and point mutations. At each generation, two most fittest members were mated to produce 8 offsprings. The six parameters that were explored were: spatial and temporal pooling ranges plus the number of neurons in a competition pool for two adjacent layers. In other words, six fixed parameters in Table (namely 6 and 1 for the pooling range and 8 for the number of

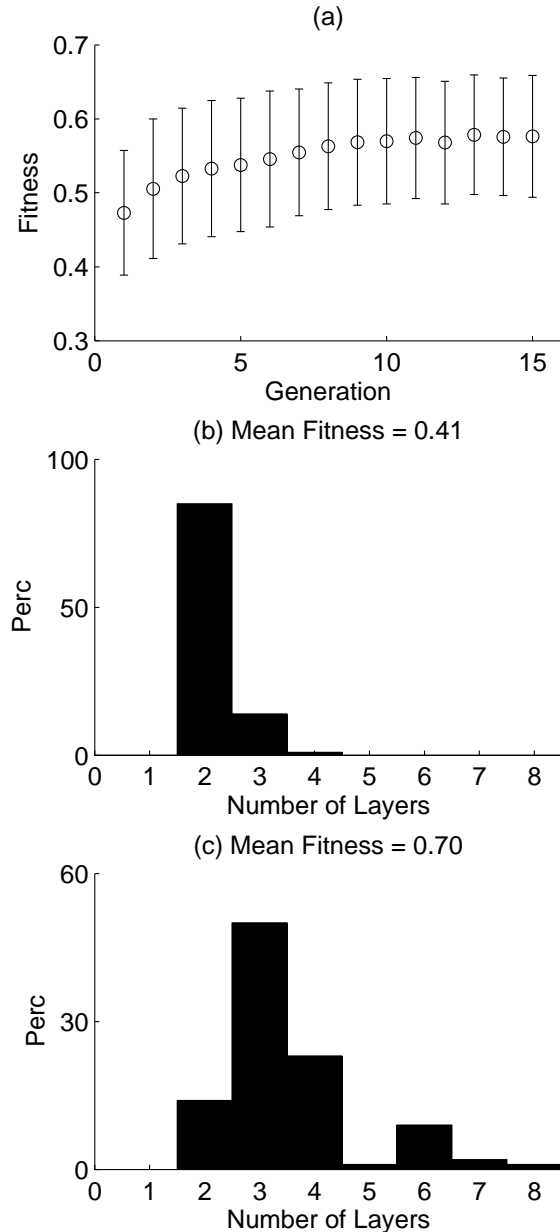


FIG. 4. Genetic algorithm for optimizing the network architecture. (a) The fitness of neural networks, as defined by their classification performance on translated hand-written digit images (chance = 0.1), increases over generations. The genetic algorithm finds that deeper networks on average perform better than the shallower ones, as shown by the histograms of the bottom (b) and the top (c) 100 networks. Optimized parameters were the spatial and temporal pooling ranges, as well as the number of neurons that share the same receptive field.

neurons in odd levels and 2, 4, and 4 in even levels) are now allowed to vary at each generation of the genetic algorithm. Fig. 4 shows that on average the neural networks in the later generations perform better than their ancestors. More importantly, when the number of hierarchical layers in the neural networks are compared, it is found that deeper networks outperform shallower ones in recognizing translated objects.

## DISCUSSION

This study has demonstrated that the proposed algorithm is capable of learning selectivity and tolerance by associating spatially or temporally co-occurring features. It follows the idea that the visual system takes advantage of the physical coherence of objects in the world [9, 13, 14, 20] and the hierarchical, local pooling is an important architectural principle of the neural system [1, 5]. As a proof of concept, this study has explored simple toy stimuli only, so testing the scalability of the algorithm with natural images to various transformation sequences [19, 20], as well as investigating the applicability of the same canonical algorithm in other sensory modalities [33], will be important.

---

\* mkouh@drew.edu

- [1] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, *Neuron* **73**, 415 (2012).
- [2] K. Fukushima, S. Miyake, and T. Ito, *IEEE Transactions on Systems, Man and Cybernetics* **13**, 826 (1983).
- [3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Proceedings of the IEEE* **86**, 2278 (1998).
- [4] M. Riesenhuber and T. Poggio, *Nature Neuroscience* **2**, 1019 (1999).
- [5] M. Kouh and T. Poggio, *Neural Computation* **20**, 1427 (2008).
- [6] A. J. Bell and T. J. Sejnowski, *Vision Res.* **23**, 3327 (1997).
- [7] B. A. Olshausen and D. J. Field, *Vision Research* **37**, 3311 (1997).
- [8] P. Földiák, *Biological Cybernetics* **64**, 165 (1990).
- [9] J. V. Stone, *Neural Computation* **8**, 1463 (1996).
- [10] D. O. Hebb, *Brain* **82**, 260 (1959).
- [11] Y. Dan and M. M. Poo, *Neuron* **44**, 23 (2004).
- [12] J. J. Hopfield, *Proc. Natl. Acad. Sci. USA* **79**, 2554 (1982).
- [13] P. Földiák, *Neural Computation* **3**, 194 (1991).
- [14] L. Wiskott and T. J. Sejnowski, *Neural Computation* **14**, 715 (2002).
- [15] D. Zoccolan, M. Kouh, T. Poggio, and J. J. DiCarlo, *Journal of Neuroscience* **27**, 12292 (2007).
- [16] T. O. Sharpee, M. Kouh, and J. H. Reynolds, *Proc. Natl. Acad. Sci. USA* **110**, 11618 (2013).
- [17] G. E. Hinton and R. R. Salakhutdinov, *Science* **313**, 504 (2006).
- [18] D. George and J. Hawkins, *PLoS Computational Biology* **5**, e1000532 (2009).
- [19] Q. V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng, *Proceedings of the International Conference on Machine Learning* (2011).
- [20] S. M. Stringer, G. Perry, E. T. Rolls, and J. Proske, *Biological cybernetics* **94**, 128 (2006).
- [21] V. B. Mountcastle, *Journal of Neurophysiology* **20**, 408 (1957).
- [22] V. B. Mountcastle, *Cereb. Cortex* **13**, 2 (2003).
- [23] T. Masquelier and S. J. Thorpe, *PLoS Computational Biology* **3**, e31 (2007).
- [24] D. Hubel and T. Wiesel, *Journal of Physiology* **195**, 215 (1968).
- [25] J. L. Gallant, C. E. Connor, S. Rakshit, J. W. Lewis, and D. C. Van Essen, *Journal of Neurophysiology* **76**, 2718 (1996).
- [26] S. V. David, B. Y. Hayden, and J. Gallant, *Journal of Neurophysiology* **96**, 3492 (2006).
- [27] A. Anzai, X. Peng, and D. C. Van Essen, *Nature neuroscience* **10**, 1313 (2007).
- [28] J. Hegde and D. C. Van Essen, *Cereb. Cortex* **17**, 1100 (2007).
- [29] T. Hirabayashi, D. Takeuchi, K. Tamura, and Y. Miyashita, *Science* **341**, 191 (2013).
- [30] N. C. Rust and J. J. DiCarlo, *Journal of Neuroscience* **30**, 12978 (2010).
- [31] T. J. Anastasio, *Tutorial on neural systems modeling* (Sinauer, 2010).
- [32] A. K. McAllister, D. C. Lo, and L. C. Katz, *Neuron* **15**, 791 (1995).
- [33] C. Kanan, *PLoS One* **8**, e54088 (2013).